Synesthesia via Direct Latent Augmentation: Bypassing the Decode-Encode Loop for Cross-Modal Distillation

Cristian Sbrolli[1], Nicolas Michel[2], Matteo Matteucci[1], and Toshihiko Yamasaki[2]

[1] Politecnico di Milano, Milan, Italy
[2] University of Tokyo, Tokyo, Japan
cristian.sbrolli@polimi.it

Abstract. While multimodal integration significantly improves computer vision models, deploying them incurs prohibitive inference costs and requires scarce, perfectly paired datasets. Recent methods address this data bottleneck by synthesizing missing modalities via generative AI, yet they introduce a severe inefficiency: the Decode-Encode Loop. Specifically, information-rich generative latents are decoded into noisy raw signals, forcing the downstream classifier to waste capacity re-encoding them. To bypass this bottleneck, we propose Direct Latent Augmentation (DLA), utilizing undecoded generative latents directly as privileged information. Furthermore, to transfer this dense knowledge to a purely visual student, we introduce Multilayer Explicit Simulated Synesthesia (MESSy). Instead of enforcing rigid representation matching, which forces the student to distort its native visual features to accommodate complex multimodal topologies, MESSy uses a predictive objective to safely internalize these physical priors. Empirical results demonstrate that our framework significantly outperforms raw data augmentation and traditional distillation. Ultimately, our approach yields highly accurate unimodal students with "synesthetic" latent structures that are inherently aligned with modalities they have never directly observed.

Keywords: Knowledge Distillation, Multimodal Learning, Generative Data Augmentation, Latent Representations

1 Introduction

Human perception is intrinsically multimodal. We understand the physical world not just through visual appearance, but through a continuous integration of sensory signals. The sight of a glass shattering is deeply coupled with its acoustic signature and its geometric fragility. Similarly, neural architectures designed to fuse multiple modalities, such as 2D images with 3D objects and audio, consistently demonstrate superior robustness and accuracy compared to unimodal

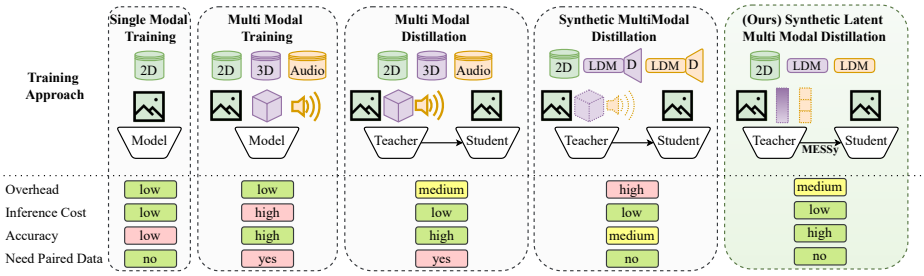

Fig. 1: Comparison of multimodal learning paradigms. Traditional multimodal networks achieve high accuracy but require scarce paired datasets and incur prohibitive inference costs. Cross-modal distillation solves the inference bottleneck but remains data-dependent. While naive synthetic distillation removes the need for paired data, it introduces a severe training overhead due to the expensive generative decode-encode loop. Our framework (right) bypasses this bottleneck entirely. By directly distilling undecoded generative latents into a purely visual student, we achieve high multimodal accuracy and efficient unimodal inference without requiring real paired datasets. Our framework reduces 3D generation time by 87.7% and audio VRAM by 23.8% relative to full decoding.

models. Additional modalities provide indeed complementary physical and semantic priors that help resolve visual ambiguities.

Despite their performance advantages, deploying multimodal networks in real-world applications presents significant challenges. Extracting and processing multiple high-dimensional data streams incurs a substantial computational cost during inference. More practically, auxiliary modalities like structured 3D geometry or clean audio are rarely available in unconstrained test environments. An effective solution to this deployment problem is the paradigm of Learning Using Privileged Information (LUPI) [42] via cross-modal knowledge distillation. By training a heavy multimodal teacher and distilling its knowledge into a lightweight, image-only student, we can achieve high performance while maintaining unimodal inference efficiency.

However, this approach suffers from a severe data dependency. While image and text pairs can be readily scraped from the web at an unprecedented scale, data for continuous physical modalities such as high-fidelity audio and 3D geometry remains inherently scarce in isolation, and prohibitively rare when strict cross-modal alignment is required.

The rapid advancement of generative artificial intelligence offers a promising mechanism to overcome this data scarcity. Modern cross-modal generators can synthesize highly realistic missing modalities directly from visual inputs or their corresponding text captions [47]. An emerging trend in representation learning exploits these models to perform synthetic multimodal augmentation, creating artificial pairs to train teacher networks [6, 17, 30]. While generating synthetic data circumvents the need for real paired datasets, current pipelines overlook a key structural feature of modern generative models. State-of-the-art generators,

such as Latent Diffusion Models (LDM), operate within compressed, semantically dense latent spaces. Forcing these latents through an expensive decoding step to produce "raw" data creates a highly inefficient *Decode-Encode Loop*. The decoder inevitably injects artifacts and noise irrelevant to classification [43], forcing the downstream multimodal classifier to waste parameter capacity re-encoding this noisy data back into a semantic vector.

We propose Direct Latent Augmentation (DLA) to bypass this inefficiency entirely. Instead of fully decoding the synthetic data, we intercept the generative process and extract the undecoded latents. We treat these latents as pristine privileged information. Because they are extracted prior to the generative decoder, they retain maximum semantic density, are free from decoding artifacts, and require a fraction of the computational cost to load and process. We fuse these heterogeneous latents with 2D image patches using a Perceiver-based latent resampler to train our Multimodal Teacher.

Transferring this dense multimodal knowledge to a standard Vision Transformer [11] student presents a secondary challenge. Existing cross-modal distillation, together with standard knowledge distillation loss, minimizes the direct Euclidean distance between the teacher and student embedding spaces. We argue this imposes a rigid geometric isomorphism that causes *capacity interference*: forcing a purely visual student to perfectly mimic a high-dimensional space dedicated to audio and 3D geometry degrades its native visual discriminative ability. To transfer this knowledge effectively, we introduce **Multilayer Explicit Simulated Synesthesia (MESSy)**. Instead of forcing the student's latent space to identically match the teacher's, MESSy employs lightweight predictive heads. The student learns to predict the teacher's auxiliary modality features from their own visual tokens. This biomimetic approach is inspired by human synesthesia, ensuring the student internalizes the sufficient statistics to recall physical properties without collapsing its native visual manifold. Eventually, we validate our framework on Imagenette, Caltech101, and ImageNet-100. Our contributions are summarized as follows:

1. We identify the inefficiencies of the Decode-Encode Loop in generative data augmentation and propose Direct Latent Augmentation (DLA), demonstrating that generative latents provide a cleaner, more efficient supervision signal than raw synthetic data.
2. We introduce Multilayer Explicit Simulated Synesthesia (MESSy) to solve capacity interference during cross-modal distillation, allowing unimodal students to safely internalize multimodal concepts.
3. We provide quantitative and qualitative evidence of Artificial Synesthesia, demonstrating that the unimodal student structurally aligns its visual latent space with unobserved physical properties and selectively leverages acoustic and geometric priors to resolve visual ambiguities.

2 Related Work

2.1 Multimodal Classification and LUPI

Fusing multiple sensory modalities, such as 2D images with audio waveforms or 3D shapes, consistently improves performance on standard classification tasks by leveraging complementary physical priors [24, 31]. Architectures employing late-fusion, cross-modal attention, or attention bottlenecks [23, 45] effectively aggregate these heterogeneous inputs to resolve visual ambiguities. However, deploying such networks in the wild remains largely impractical. Processing high-dimensional, multi-sensory data incurs substantial inference costs and fundamentally relies on the availability of multimodal data at test time, which is frequently absent in real-world applications.

To bridge the gap between the superior accuracy of multimodal training and the efficiency of unimodal inference, researchers rely on the Learning Using Privileged Information (LUPI) paradigm [42]. LUPI operates through cross-modal knowledge distillation [19], where a multimodal teacher guides a unimodal student. The pioneering work of Gupta et al. [16] established RGB-to-depth and RGB-to-flow supervision transfer as a cross-modal distillation paradigm, directly inspiring subsequent LUPI pipelines [14, 20, 30, 46]. Despite their success, these methods share a fundamental limitation: they require training datasets of perfectly aligned, real-world multi-sensory triplets. Curating such physically grounded paired data is expensive and difficult to scale, severely limiting the broad application of traditional cross-modal distillation for standard vision tasks.

2.2 Generative Augmentation and the Decode-Encode Bottleneck

The rapid maturation of generative AI provides a solution to data scarcity. Generative models are increasingly used for synthetic data augmentation, creating novel views, depth maps, or text captions to enhance unimodal classifiers [3, 12, 18].

Closely related to our objective is the recent concept of Learning Using Generated Privileged Information (LUGPI) [30]. LUGPI demonstrates that generating the image modality from text prompts via diffusion and treating it as privileged information improves student classification accuracy. However, their scope remains confined to the highly abundant text and image domains, neglecting the scarce physical modalities, such as 3D structure and audio, that are essential human-like understanding. Crucially, pipelines such as LUGPI operate exclusively in the raw data domain, fully decoding generated signals into human-perceptible formats, like high-resolution images.

Feature-level augmentation within a single visual encoder has also been explored to improve robustness to affine transformations [36]. In contrast, our DLA framework generates auxiliary-modality latents from frozen cross-modal generators as privileged information, operating across modality boundaries rather than perturbing the model's own feature space.

We argue that this introduces a significant computational and representational inefficiency. Because state-of-the-art generators already perform dense semantic compression in a latent space [25, 34], pushing these representations through a final decoder merely to accommodate human perception is counterproductive. This decoding step injects generative artifacts, phase noise, and high-frequency details that act as nuisance variables for classification. Consequently, the downstream multimodal teacher wastes parameter capacity re-encoding this noisy raw data. Our Direct Latent Augmentation (DLA) framework bypasses this decode-encode bottleneck entirely, utilizing the undecoded generative latents directly to provide a cleaner and computationally cheaper supervision signal.

2.3 Feature-Level Distillation and Capacity Interference

Knowledge distillation traditionally aligns logits [19]; FitNets [35] extended this to projector-based hidden-layer alignment, with multi-layer targets stabilizing optimization [37, 41]. Asymmetric predictor heads, where a student predicts frozen teacher representations, underpin BYOL [15], data2vec [4, 5], and their cross-modal extension AV-data2vec [27], the closest architectural predecessor to MESSy. JEPA variants [2, 22] similarly predict abstract teacher embeddings rather than raw pixels. CRD [40] shows that relaxing rigid L_2 alignment improves cross-modal transfer; Zhao et al. [48] independently formalize the capacity-interference problem and propose a trainable projector as remedy.

Cross-modal LUPI pipelines standardly penalize the L_2 distance between the student's unimodal and the teacher's multimodal embedding [14,30,46], inducing *capacity interference* (Sec. 3.4). MESSy adopts the predictor-head pattern of the above methods but targets pooled summaries of *unobserved synthetic auxiliary-modality* tokens, distinct from self-features [2, 5], the teacher CLS [30], or same-modality perturbations [36] preserving the student's native visual manifold.

3 Proposed Method

3.1 The Advantage of Latents and Multimodal Synergy

Recent advancements in generative AI, particularly Latent Diffusion Models (LDMs) and Flow Matching [28, 34], have established a powerful paradigm for high-fidelity data synthesis. These models operate fundamentally in two stages. First, a generative process iteratively transforms initial noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ into a structured, semantically dense representation Z within a compressed latent space, guided by a conditioning signal c. Second, a pre-trained decoder $\mathcal{D}$ maps this latent vector into the high-dimensional raw data space, producing the human-perceptible signal $\hat{X} = \mathcal{D}(Z)$ (such as an audio waveform or a 3D mesh).

In the context of multimodal data augmentation, the goal is to provide a classifier with auxiliary physical signals that correlate with the target label Y. Standard synthetic augmentation pipelines utilize the fully decoded raw data $\hat{X}$. To process this data, the downstream multimodal classifier must employ a modality-specific encoder $\mathcal{E}$, yielding a feature representation $F = \mathcal{E}(\hat{X})$.

This establishes a deterministic Markov chain from the generated latent to the final classification features: $Z \xrightarrow{\mathcal{D}} \hat{X} \xrightarrow{\mathcal{E}} F$. By the Data Processing Inequality [8], the mutual information I between the label Y and any representation along this chain cannot increase:

$$I(Y;Z) \geq I(Y;\hat{X}) \geq I(Y;F) \tag{1}$$

This inequality exposes a fundamental inefficiency in current synthetic augmentation pipelines. The decoder $\mathcal{D}$ is optimized for human perception, not machine classification. To bridge the gap between the compressed latent space and the raw data manifold, $\mathcal{D}$ injects high-frequency structural details, phase variations, and unavoidable generative artifacts. From the perspective of the classifier, this injected information acts as a noisy channel that degrades the Signal-to-Noise Ratio (SNR) of the underlying semantics. Consequently, the downstream encoder $\mathcal{E}$ must waste parameter capacity attempting to filter out this perceptual noise to recover the semantic concepts already present in Z.

Furthermore, our framework allows us to efficiently scale the number of conditioning modalities without amplifying this encoding noise. From an information-theoretic perspective, integrating multiple latent modalities, such as visual (Z_v), acoustic (Z_a), and geometric (Z_g) representations, strictly expands the available semantic signal. By the chain rule of mutual information, the joint information shared with the target label Y is:

$$I(Y;Z_v,Z_a,Z_g) = I(Y;Z_v) + I(Y;Z_a|Z_v) + I(Y;Z_g|Z_v,Z_a) \tag{2}$$

Because conditional mutual information is inherently non-negative:

$$I(Y;Z_v,Z_a,Z_g) \geq I(Y;Z_v) \tag{3}$$

Thus, each additional latent space has the potential to provide orthogonal, non-redundant cues, such as structural morphology from 3D latents or unique acoustic signatures, that resolve ambiguities present in the visual signal alone.

By intercepting the generative pipeline and utilizing the undecoded latents directly as inputs to the multimodal classifier, our framework theoretically guarantees a higher upper bound on I. This provides a cleaner, dense supervision signal that maximizes multimodal synergy, while drastically reducing the computational overhead of decoding and re-encoding continuous physical modalities.

3.2 Stage 1: Synthetic Latent Augmentation

The first stage of our framework constructs an offline, multimodal latent dataset from a unimodal image corpus (Phase 1, Fig. 2). For a given RGB image query X_v, we extract semantically corresponding generative latents for 3D geometry and audio without ever synthesizing the final raw signals.

For the 3D modality, we employ the pre-trained Hunyuan3D-2 Latent DiT [39] to obtain the X_v image-conditioned continuous 3D latents, $Z_{3D} \in \mathbb{R}^{N_{3D} \times D_{3D}}$.

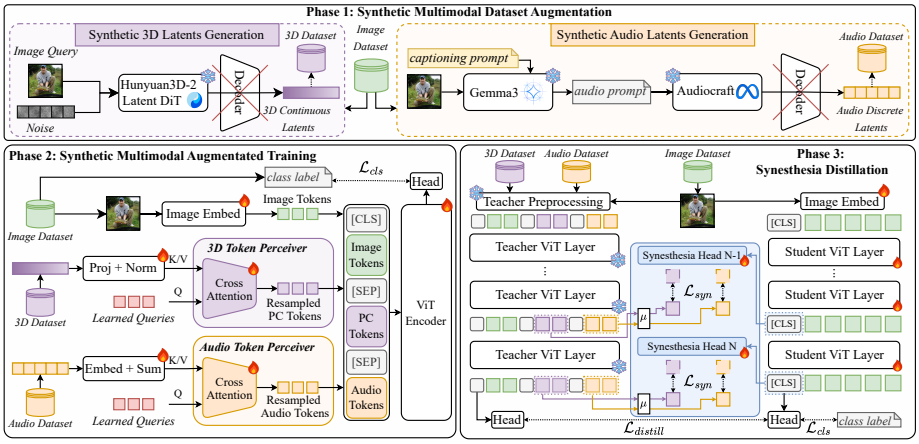

Fig. 2: Phase 1 (Synthetic Latent Augmentation): Given an image query, we extract undecoded continuous 3D latents and discrete audio codes directly from generative models, deliberately bypassing the noisy and computationally expensive Decode-Encode Loop. **Phase 2 (Multimodal Teacher Training):** A Multimodal Teacher is trained to classify these augmented triplets. We fuse standard image tokens with the heterogeneous generative latents using modality-specific Perceiver cross-attention blocks. **Phase 3 (MESSy Distillation):** We transfer this multimodal knowledge to a lightweight, purely visual student. Using our MESSy objective, lightweight auxiliary heads predict the teacher's multimodal features across multiple intermediate layers. This allows the student to learn the physical priors avoiding capacity interference.

Crucially, the 3D latent-to-mesh decoder is bypassed, saving significant computational time and preserving semantic density.

For the audio modality, direct image-to-audio generative models remain scarce and lack the robust, zero-shot generalization capabilities of their text-conditioned counterparts. To bridge this modality gap, we employ Gemma3 [38] to process the image query and generate a descriptive captioning prompt of the likely acoustic events in the scene. To strictly prevent label leakage, this prompting process is carefully designed: the model is never provided with the ground-truth class name, and we implement an automatic filtering mechanism that regenerates the text if the generated caption inadvertently contains the class name or any of its WordNet synonyms (full prompt and pipeline details are provided in the supplementary material). We then input this audio prompt into Audiocraft [7]. Because audio generators operate using discrete latent spaces quantized by neural audio codecs, we intercept the generation process immediately after the autoregressive transformer samples the acoustic codebook, extracting the undecoded Audio Discrete Latents $Z_{aud} \in \mathbb{Z}^{N_{aud}}$. We deliberately extract and save the undecoded discrete latent codes Z_{aud} rather than the generator's internal continuous embeddings. This not only maximizes offline storage and I/O efficiency, but also allows our downstream Teacher network to learn a task-specific embedding lookup that prioritizes semantic classification over acoustic waveform reconstruction.

Both Z_{3D} and Z_{aud} are saved to disk, creating a highly compressed, information-dense multimodal dataset while completely avoiding the Decode-Encode Loop. We report full per-modality generation details in the supplementary material.

3.3 Stage 2: Synthetic Multimodal Augmented Training

In Phase 2, we train a multimodal teacher network, $\mathcal{T}$, to classify the augmented triplets (X_v, Z_{3D}, Z_{aud}). Fusing these modalities presents a dimensional challenge: the visual input consists of standard spatial patches, the 3D input consists of continuous latent sequences of variable length, and the audio input consists of discrete codebook tokens.

To project these heterogeneous inputs into a unified representation space, we design a modality-agnostic fusion architecture utilizing Perceiver-inspired [1, 23] cross-attention blocks. The visual image X_v is processed through a standard patch embedding layer to produce Image Tokens E_v. For the 3D modality, the continuous latents Z_{3D} are projected and normalized, then passed as Keys and Values to a 3D Token Perceiver. By cross-attending with a fixed number of Learned Queries, the Perceiver compresses the variable-length latents into a fixed budget of Resampled 3D Tokens E_{3D}.

Similarly, the discrete audio codes Z_{aud} are passed through a learned embedding lookup, summed, and fed as Keys and Values to an Audio Token Perceiver. This outputs a fixed budget of Resampled Audio Tokens E_{aud}.

The final input sequence to the Teacher's ViT Encoder is constructed by concatenating the modality tokens, explicitly bounded by learnable separator tokens ([SEP]), alongside a prepended classification token ([CLS]):

$$E_{input} = [[\texttt{CLS}], E_v, [\texttt{SEP}_1], E_{3D}, [\texttt{SEP}_2], E_{aud}] \tag{4}$$

The Multimodal Teacher $\mathcal{T}$ processes this joint sequence and is optimized using a standard cross-entropy loss, $\mathcal{L}_{\text{cls}}$, against the ground-truth class labels.

3.4 Stage 3: Distillation via Multilayer Explicit Simulated Synesthesia (MESSy)

In the final phase, our objective is to transfer the robust multimodal representations of the frozen teacher $\mathcal{T}$ into a lightweight, unimodal student network $\mathcal{S}$, which receives only the standard RGB images as input. Along with standard logits distillation, recent works using LUPI and LUGPI frameworks achieve this by an additional loss term enforcing a strict geometric isomorphism, minimizing the direct L_2 distance between the student's visual embedding and the teacher's multimodal embedding. However, we argue this is fundamentally flawed for asymmetric cross-modal distillation. The teacher's embedding resides in a joint manifold shaped by visual, acoustic, and geometric features, whereas the student is constrained by the information capacity of a purely visual input. Forcing a lower-capacity unimodal embedding to perfectly mimic this high-dimensional topology

induces *capacity interference*: a phenomenon where the network struggles to simultaneously optimize for fundamentally conflicting representations, leading to destructive gradient conflict [44]. The student is forced to distort its optimal visual decision boundaries to geometrically accommodate the missing modalities, which actively degrades its primary discriminative capabilities.

To resolve this capacity bottleneck, we propose Multilayer Explicit Simulated Synesthesia (MESSy), illustrated in Phase 3 of Figure 2. Rather than enforcing rigid L_2 spatial equality $||e^S - e^T||_2^2$ between the visual student embedding e^S and the multimodal teacher embedding e^T, MESSy requires the student to encode only the sufficient statistics necessary to reconstruct the auxiliary modalities, thereby inducing "Simulated Synesthesia". To ensure computational efficiency and avoid the overhead of sequence-to-sequence prediction, we do not force the student to reconstruct the full sequence of synthetic tokens. Instead, let $e_{aud}^{T,l}$ and $e_{3D}^{T,l}$ represent the global, average-pooled summary vectors of the acoustic and geometric tokens extracted from the teacher at layer l. We attach lightweight, non-linear projection heads H_{aud} and H_{3D} (implemented as 2-layer MLPs) to the student's corresponding classification tokens $e_{\text{CLS}}^{S,l}$. The MESSy objective is defined as the MSE over the deepest K layers of the Vision Transformer:

$$\mathcal{L}_{\text{MESSy}} = \sum_{l=L-K}^{L} \mathbb{E}\left[||H_{aud}(e_{\text{CLS}}^{S,l}) - e_{aud}^{T,l}||_2^2 + ||H_{3D}(e_{\text{CLS}}^{S,l}) - e_{3D}^{T,l}||_2^2\right] \quad (5)$$

From an Information Bottleneck perspective, this predictive objective maximizes the mutual information $I(e_{\text{CLS}}^S; e_{aud}^T, e_{3D}^T)$ by utilizing the projection heads to bridge the spatial gap between the manifolds, safely bypassing capacity collapse. We apply this to multiple layers following recent advancements in transformer distillation showing that multi-layer supervision stabilizes the optimization trajectory of self-attention blocks [37, 41]. By applying MESSy across the final K layers, we provide useful gradient signals without interfering with early-stage visual feature extraction (we ablate this number in the supplementary material).

During inference, the auxiliary heads H are entirely discarded, leaving the student with an enriched visual latent space and zero additional computational overhead. The total training objective combines standard cross-entropy ($\mathcal{L}_{\text{cls}}$) against the hard labels, classical logit-level knowledge distillation ($\mathcal{L}_{\text{distill}}$), and our explicit synesthesia loss ($\mathcal{L}_{\text{MESSy}}$):

$$\mathcal{L}_{\text{Total}} = \alpha\mathcal{L}_{\text{cls}} + (1-\alpha)\mathcal{L}_{\text{distill}} + \beta\mathcal{L}_{\text{MESSy}} \quad (6)$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on Imagenette2-320 [21], Caltech101 [26], and ImageNet-100 [10,33]. Focusing on these benchmarks allows us to rigorously ablate modality combinations and perform expensive raw-vs-latent generative comparisons

within a feasible computational budget. To verify that our method scales beyond these benchmarks, we also evaluate our best DLA+MESSy configuration on the full ImageNet-1K dataset [10] (Sec. 4.6). We execute 3 independent runs per experiment, reporting averages and standard deviations ($\pm$) in tables, and averages in plots for clarity.

Generative Pipeline. For 3D latent extraction, we utilize the pre-trained `Hunyuan3D-2-mini` Latent Diffusion Model (FP16 variant). We extract the continuous image-conditioned spatial latents directly from the DiT flow-matching pipeline, bypassing the final mesh decoder. For audio latent extraction, we first prompt the `Gemma-3-12B-IT` large language model with the visual image to generate a short, descriptive acoustic caption. This caption conditions `AudioGen-Medium`. We intercept the generation process at the autoregressive transformer output, extracting the discrete tokens from the 4 parallel acoustic codebooks prior to EnCodec waveform decoding.

Architecture and Training. Both our multimodal teacher ($\mathcal{T}$) and unimodal student ($\mathcal{S}$) are based on the standard Vision Transformer (ViT-B/16) architecture, ensuring fair capacity comparisons. For the Teacher, we introduce modality-specific Latent Resamplers using Perceiver-style cross-attention with 8 heads. These resamplers compress the variable-length generative latents into 100 tokens for 3D geometry and 100 tokens for audio. During Teacher training, we apply an aggressive modality dropout rate of $p = 0.8$ to the acoustic and geometric tokens. We find this to be crucial to prevent the model from over-relying on the synthetic modalities (detailed in the supplementary material).

Distillation Details. The Teacher is frozen after 50 epochs of training. The Student is then trained from scratch for 50 epochs using AdamW [29] optimization and a cosine annealing learning rate scheduler. For the distillation objective, we set the standard Knowledge Distillation temperature to $T = 2.0$ and the Cross-Entropy weight to $\alpha = 0.25$. For the MESSy objective, we extract features from the last $K = 3$ layers (ablated in the supplementary material) and we set the loss weight to $\beta = 1.0$.

4.2 Multimodal Teacher Performance: The Latent Advantage

Before evaluating downstream accuracy, we quantify the computational savings of this bypass during the offline dataset augmentation phase. Bypassing the expensive mesh decoder for 3D generation reduces inference time by 87.7% and saves over 600 MB of peak VRAM. Similarly, extracting discrete audio tokens prior to waveform decoding yields a 20.6% reduction in generation time and cuts peak VRAM usage by 23.8% (nearly 2.8 GB). Full benchmarking details are provided in the supplementary material.

Figure 3 and Tab. 1 compare multimodal teachers accuracies when augmented with raw vs latent synthetic data. For the raw synthetic data encoding we use a PointNet++ style local feature tokenizer, while we convert WAV to Mel-spectrograms and then embed them via 2D convolution. Training the Teacher directly on generative latents significantly outperforms training on fully decoded raw synthetic data across all benchmarks. For instance, on Caltech101,

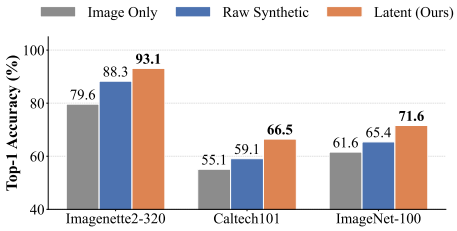

Fig. 3: Representation efficiency of the Multimodal Teacher. DLA consistently outperforms fully decoded raw synthetic data across all benchmarks.

Table 1: Multimodal Teacher validation accuracies. We compare the Image-Only baseline, fully decoded Raw synthetic data, Raw data with the same Perceiver resampler architecture, and our undecoded Generative Latents.

Method	Imagenette2	Caltech101	ImageNet-100
Image Only	79.63 ± 0.32	55.10 ± 0.27	61.57 ± 0.53
Raw	88.28 ± 0.65	59.08 ± 1.96	65.40 ± 0.90
Raw+Perc.	88.92 ± 0.43	60.21 ± 1.12	66.13 ± 0.77
Latents	**93.10** ± 0.19	**66.47** ± 1.42	**71.59** ± 0.45

the latent teacher achieves an absolute improvement of +7.39% over the raw synthetic teacher and +11.37% over the Image-only baseline, with similarly substantial gains observed on ImageNet-100 and Imagenette2-320. These results empirically validate our analysis (Sec. 3.1): generative decoding injects perceptual noise that degrades the semantic Signal-to-Noise Ratio, whereas undecoded latents provide dense, highly discriminative privileged information. To rule out that this gain arises from a difference in encoder architecture rather than representation quality, we also train a Raw teacher that feeds decoded modalities through the same Perceiver resampler used for latents (*Raw+Perc.* in Tab. 1). The Perceiver resampler accounts for at most 1.13 pp (≈14% of the gap on average), confirming that the dominant factor is the undecoded representation itself, in line with the Data Processing Inequality.

Furthermore, we perform a study on the relative contribution of each modality Fig. 4 and Tab. 2. We observe that 3D geometry and audio provide orthogonal, complementary supervision. On ImageNet-100, the Image+3D (70.03%) and Image+Audio (65.76%) combinations both independently improve upon the Image-Only baseline (61.57%). Fusing all modalities (the "Full" configuration) consistently yields the highest upper bound performance (71.59%).

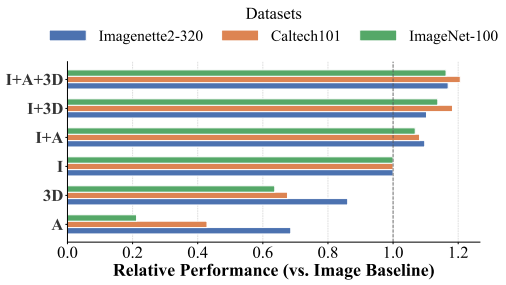

Fig. 4: Modality contribution analysis.

Crucially, we evaluate the networks trained solely on the isolated synthetic modalities to ensure the model is not simply exploiting a generative shortcut. We find that the synthetic modalities perform poorly on their own, falling far below the real Image-Only baseline. The Audio-Only baseline achieves 13.07% on ImageNet-100, well above the 1% random chance for 100 classes, which confirms

Table 2: Performance comparison across different modality combinations. We report the average validation accuracy and standard deviation (±) across 3 runs, alongside the absolute performance delta (Δ) relative to the Image-Only baseline.

Modalities	Imagenette2		Caltech101		ImageNet-100	
	Accuracy	Δ	Accuracy	Δ	Accuracy	Δ
Image Only	79.63 ± 0.32	–	55.10 ± 0.27	–	61.57 ± 0.53	–
Audio Only	54.61 ± 0.85	−25.02	23.61 ± 1.10	−31.49	13.07 ± 0.54	−48.50
3D Only	68.53 ± 0.43	−11.10	37.23 ± 2.84	−17.87	39.20 ± 0.73	−22.37
Image + Audio	87.37 ± 0.24	+7.74	59.58 ± 0.83	+4.48	65.76 ± 0.04	+4.19
Image + 3D	87.80 ± 0.15	+8.17	65.15 ± 0.24	+10.05	70.03 ± 0.53	+8.46
Image + 3D + Audio	**93.10** ± 0.19	**+13.47**	**66.47** ± 1.42	**+11.37**	**71.59** ± 0.45	**+10.02**

that the acoustic latents do carry class-relevant signal, yet far below the 61.57% Image-Only accuracy, ruling out any trivial label leakage from the generation pipeline. The 3D-Only baseline (39.20%) is stronger, reflecting the richer spatial structure encoded by the Hunyuan3D latents. Both modalities therefore act as complementary physical priors that sharpen the visual encoder rather than as shortcuts, validating the integrity of our label-leakage prevention pipeline.

4.3 Synesthesia Distillation Efficacy

Having established a robust multimodal upper bound, we evaluate the efficacy of transferring this knowledge to a unimodal, image-only Student ($\mathcal{S}$) (Fig. 5 and Tab. 3). Our MESSy Student achieves the best unimodal inference performance, outperforming the baseline by +8.42% on Caltech101 and +6.97% on ImageNet-100, effectively bridging the gap to the synthetic upper bound.

We compare MESSy against KD and LUGPI, all using the same DLA latent teacher (LUGPI applies L_2 matching to our synthetic representations, not a raw-data teacher). KD yields only marginal gains: logit-level matching cannot capture the geometric and acoustic semantics embedded in the teacher's intermediate representations. LUGPI improves substantially but is limited by the rigid L_2 alignment it imposes; MESSy's predictive heads relax this constraint, adding +3.40% over LUGPI on Caltech101 and +2.84% on ImageNet-100 (Sec. 3.4).

Capacity interference evidence. We freeze both IN-100 students and evaluate them on a CIFAR-100 linear probe, a purely visual task seen by neither student during training. LUGPI scores 42.4% vs. MESSy's 48.1% (+5.7 pp). Since both students share the same teacher and training images, the gap can only be attributed to the distillation objective: L_2 matching degrades visual discriminability to accommodate the multimodal teacher topology, while MESSy's predictive heads preserve it.

MESSy target ablation and DLA isolation. Tab. 4 ablates two factors on ImageNet-100. A CLS-Pred variant uses the same MLP heads but targets teacher CLS tokens: switching from L_2 to a projection head adds +1.34 pp; switching the prediction target to auxiliary-modality summaries adds +1.50 pp, showing the gain is not merely architectural. Replacing the DLA teacher with a raw-data

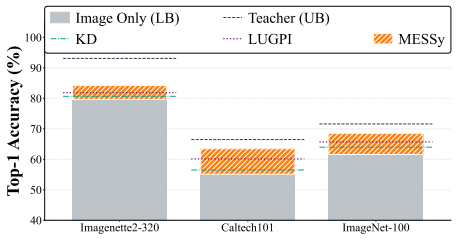

Fig. 5: Our approach (MESSy) against the RGB-only Lower Bound, KD, LUGPI, and the Synthetic Upper Bound (Teacher UB).

Method	Imagenette2	Caltech101	ImageNet-100
Image	79.63 ± 0.32	55.10 ± 0.27	61.57 ± 0.53
Synth. UB	93.10 ± 0.19	66.47 ± 1.42	71.59 ± 0.45
KD	80.66 ± 0.20	56.49 ± 0.42	63.97 ± 0.59
LUGPI	81.85 ± 0.51	60.12 ± 0.76	65.70 ± 0.68
MESSy	**84.23** ± 0.40	**63.52** ± 0.91	**68.54** ± 0.83

Table 3: Top-1 Accuracy (%) comparison of unimodal distillation paradigms. *Synth. UB* is the DLA latent teacher trained on RGB + synthetic modalities. All student models evaluate **purely on visual inputs** at inference time.

teacher costs 3.39 pp under LUGPI and 4.51 pp under MESSy, indicating the latent-space advantage carries through to the student.

4.4 Emergent Synesthesia

To confirm emergent synesthesia, the student's visual latent space must reflect the topology of unobserved modalities. We cluster 3D and audio independently into ground-truth topological spaces and measure their NMI against the model's visual embedding clusters. Two feature types per modality span different abstraction levels: *Hand-Crafted (HC)*, covering low-level acoustics (MFCCs [9], spectral centroids) and covariance-based geometry; and *Deep*, using CLAP [13] for audio and PointNet [32] for 3D. Cluster counts scale with class vocabulary (50 for Imagenette2, 150 otherwise; details in supplementary).

As shown in Fig. 6, the baseline aligns poorly with both modality spaces. LUGPI improves alignment on the Deep semantic dimension (e.g. Audio Deep NMI +0.111 on ImageNet-100) but not on the physical HC space (+0.026): L_2 matching transfers high-level semantics but cannot imprint low-level physical topology. MESSy yields larger and more consistent gains across both dimensions (Audio Deep +0.154, Audio HC +0.135). The HC improvement is particularly telling: these features encode raw acoustic statistics and geometric covariance with no semantic content, so the alignment cannot arise from label co-occurrence and is evidence of genuine structural synesthesia.

4.5 Qualitative Analysis: Disentangling Modality-Driven Priors

To understand the mechanisms driving our performance gains, we perform a qualitative class-level analysis. We compare the Image baseline against students distilled from bimodal and Full teachers. While the Full student achieves the highest overall accuracy, analyzing the bimodal models isolates the impact of specific physical priors (Fig. 7).

Acoustically Driven Improvements. The left column of Fig. 7 shows classes where the Image+Audio student yields the largest gains. These classes (e.g.,

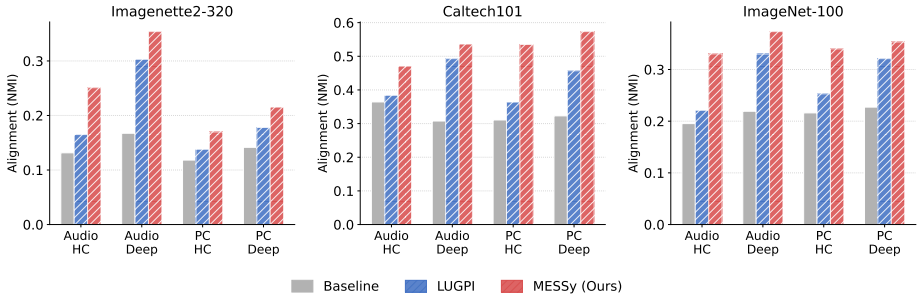

Fig. 6: Emergent alignment between the models' visual latent spaces and the unseen modalities. We compare the Image-Only baseline, the LUGPI-distilled student, and our MESSy student across both low-level physical (HC) and semantic (Deep) feature clusters. LUGPI improves Deep semantic alignment but barely affects low-level HC features. MESSy yields consistently larger and more balanced gains across both spaces.

Table 4: Ablations on ImageNet-100. *Left:* distillation objective. *Right:* DLA vs. raw teacher per objective.

Objective	Top-1	Teacher	Obj.	Top-1
LUGPI	65.70 ± 0.68	Raw	LUGPI	∼62.31
CLS-Pred	67.04 ± 0.61	Raw	MESSy	∼64.03
MESSy	$\mathbf{68.54 \pm 0.83}$	DLA	LUGPI	65.70 ± 0.68
		DLA	**MESSy**	$\mathbf{68.54 \pm 0.83}$

Table 5: Scalability on ImageNet-1K.

Method	Top-1	Δ
Image-Only	64.17	–
Synth. UB	71.32	+7.15
MESSy	**69.05**	**+4.88**

Chain Saw, *French Horn*, *Mandolin*) are defined by highly distinct acoustic signatures. By distilling acoustic latents via MESSy, the visual encoder leverages simulated sound profiles to resolve visual ambiguities, such as distinguishing a mandolin from similarly shaped stringed instruments.

Geometrically Driven Improvements. The right column highlights classes benefiting most from the Image+3D student, predominantly animals with unique morphological traits (e.g., the *Gerenuk*'s long neck or the *Hammerhead*'s cephalofoil). Distillation from 3D latents forces the student to internalize a robust spatial prior, making its vision-only representations significantly more effective.

These results empirically validate Simulated Synesthesia: the unimodal student successfully internalizes intuitive acoustic and geometric priors that align with human semantic understanding.

4.6 Scalability to ImageNet-1K

To verify that DLA+MESSy scales beyond the smaller benchmarks, we run the same pipeline on the full ImageNet-1K (ILSVRC-2012) [10], processing all 1.28M training images. The DLA teacher reaches 71.32% (+7.15% over image-only), and the MESSy student reaches 69.05% (+4.88%) evaluating on RGB only. The student recovers roughly 68% of the teacher gap, consistent with results on ImageNet-100, suggesting the framework scales predictably with dataset size.

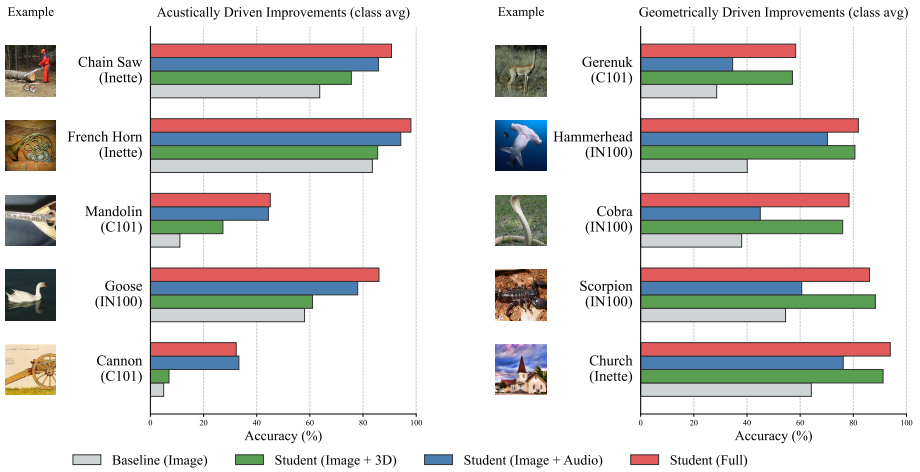

Fig. 7: Qualitative analysis of class-level accuracy improvements driven by explicitly simulated synesthesia. We compare the Image-Only baseline against students distilled from bimodal teachers (Image+3D and Image+Audio) and the Full teacher (Image+Audio+3D). **Left:** Classes where the simulated acoustic prior drives the most significant gains, typically characterized by distinct sound profiles (e.g., instruments). **Right:** Classes where the geometric prior dominates the improvement, predominantly consisting of species and objects with complex, distinctive 3D morphological traits.

5 Conclusion

We introduced DLA to bypass the inefficient decode-encode loop of generative multimodal training, utilizing pristine latents directly as privileged information. To seamlessly transfer this dense knowledge, our MESSy distillation objective allows a purely visual student to internalize acoustic and geometric priors without suffering from capacity interference. While our current framework relies on frozen, off-the-shelf generators, limiting task-specific latent co-adaptation, the empirical benefits are significant. Ultimately, our approach demonstrates that unimodal networks can achieve artificial synesthesia, successfully hallucinating multimodal properties to resolve visual ambiguities, likewise to human inherent multimodal understanding and reasoning.

Acknowledgments

This paper is supported by the PNRR-PE-AI FAIR project funded by the NextGeneration EU program. This work was partially financially supported by JST ASPIRE Program, Japan, Grant Number JPMJAP2303. This work was partially supported by the JSPS Postdoctoral Fellowship for Research in Japan (Fellowship ID: P24752).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y.S., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 23716–23736 (2022) 8
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023) 5
3. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. Transactions on Machine Learning Research (2023) 4
4. Baevski, A., Babu, A., Hsu, W.N., Auli, M.: Efficient self-supervised learning with contextualized target representations for vision, speech and language. In: International Conference on Machine Learning. pp. 1416–1429 (2023) 5
5. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning. pp. 1298–1312 (2022) 5
6. Brusini, L., Sbrolli, C., Lomurno, E., Yamasaki, T., Matteucci, M.: Polygen: Fully synthetic vision-language training via multi-generator ensembles. arXiv preprint arXiv:2602.01370 (2026) 2
7. Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., Défossez, A.: Simple and controllable music generation. In: Advances in Neural Information Processing Systems (2023) 7
8. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999) 6
9. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4), 357–366 (1980) 13
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 9, 10, 14
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 3
12. El Banani, M., Desai, K., Johnson, J.: Learning visual representations via language-guided sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19208–19220 (2023) 4
13. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap: Learning audio concepts from natural language supervision. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2023) 13
14. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: European Conference on Computer Vision. pp. 103–118 (2018) 4, 5
15. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z.D., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu,

K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020) 5
16. Gupta, S., Hoffman, J., Malik, J.: Cross-modal distillation for supervision transfer. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2827–2836 (2016) 4
17. Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: Synthclip: Are we ready for a fully synthetic clip training? arXiv preprint arXiv:2402.01832 (2024) 2
18. He, Ruifei Capsand Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? In: International Conference on Learning Representations (2023) 4
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015) 4, 5
20. Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., Darrell, T.: Cross-modal adaptation for rgb-d detection. In: IEEE International Conference on Robotics and Automation. pp. 5032–5039 (2016) 4
21. Howard, J.: Imagenette: A smaller subset of 10 easily classified classes from imagenet. `https://github.com/fastai/imagenette` (2019), accessed: June 30, 2026 9
22. Hu, N., Cheng, H., Xie, Y., Li, S., Zhu, J.: 3d-jepa: A joint embedding predictive architecture for 3d self-supervised representation learning. arXiv preprint arXiv:2409.15803 (2024) 5
23. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International Conference on Machine Learning. pp. 4651–4664 (2021) 4, 8
24. Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 (2019) 4
25. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., Adi, Y.: Audiogen: Textually guided audio generation. In: International Conference on Learning Representations (2023) 5
26. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4), 594–611 (2006) 9
27. Lian, J., Baevski, A., Hsu, W.N., Auli, M.: Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. In: IEEE Automatic Speech Recognition and Understanding Workshop. pp. 1–8 (2023) 5
28. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023) 5
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
30. Menadil, R.E., Georgescu, M.I., Ionescu, R.T.: Learning using generated privileged information by text-to-image diffusion models. In: International Conference on Pattern Recognition. pp. 423–438 (2024) 2, 4, 5
31. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: Advances in Neural Information Processing Systems. vol. 34, pp. 14200–14213 (2021) 4

32. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017) 13
33. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017) 9
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022) 5
35. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (2015) 5
36. Sandru, A., Georgescu, M.I., Ionescu, R.T.: Feature-level augmentation to improve robustness of deep neural networks to affine transformations. In: European Conference on Computer Vision. pp. 332–341 (2022) 4, 5
37. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for bert model compression. In: Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. pp. 4323–4332 (2019) 5, 9
38. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac,

J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025) 7
39. Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025) 6
40. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020) 5
41. Touvron, H., Cord, M., Douze, M.v., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357 (2021) 5, 9
42. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural Networks **22**(5-6), 544–557 (2009) 2, 4
43. Wang, M., Michel, N., Mao, J., Yamasaki, T.: Dealing with synthetic data contamination in online continual learning. Advances in Neural Information Processing Systems **37**, 51284–51311 (2024) 3
44. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12695–12705 (2020) 9
45. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020) 4
46. Yuan, S., Stenger, B., Kim, T.K.: Rgb-based 3d hand pose estimation via privileged learning with depth images. arXiv:1811.07376 (2018) 4, 5
47. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C., Xing, E.: Multimodal image synthesis and editing: The generative ai era. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15098–15119 (2023) 2
48. Zhao, C., Jin, Y., Song, Z., Chen, H., Miao, D., Hu, G.: Cross-modal distillation for widely differing modalities. arXiv preprint arXiv:2507.16296 (2025) 5